On Web-scale Reasoning

Urbani, J.

2013

Samenvatting

Over Web-scale Reasoning

Het semantische web is een uitbreiding van het huidige wereldwijde web. In het semantische web kan de betekenis van informatie door machines geïnterpreteerd worden en de data worden opgeslagen als een verzameling van subject-predicaat-object verbanden.

Op dit moment zijn er miljarden van dergelijke verbanden publiek beschikbaar op het web. Deze verbanden beschrijven een zeer breed scala van informatie: van biomedische informatie tot informatie afkomstig van regeringen. Hierbij worden URIs vaak gebruikt om concepten ondubbelzinnig te kunnen identificeren en hergebruik in gedistribueerde omgevingen, zoals het wereldwijde web, te stimuleren.

Een van de voordelen van het opslaan van informatie met gebruik van semantische web technologieën is dat computers over de data kunnen redeneren en nieuwe informatie kunnen afleiden. Dit proces, ook wel reasoning genoemd, wordt in toenemende mate uitdagender, vanwege de exponentiële groei van beschikbare data op het web. In het begin van 2009 werd het aantal van dergelijke verbanden op het semantische web geschat op 4,4 miljard. Een jaar later is de grootte van het web verdrievoudigd naar 12 miljard verbanden en de huidige trend wijst er op dat een dergelijke groei nog steeds plaatsvindt.

De onderzoeksvraag die in dit proefschrift beantwoord wordt, is: Hoe kunnen we door de inzet van reasoning op een parallel en gedistribueerd systeem de resultaten van zoekopdrachten over zeer grote hoeveelheden data verrijken?

Het proefschrift bestaat uit twee delen: in het eerste deel vindt reasoning plaats door het toepassen van een verzameling van regels over de gehele invoerdata, teneinde elke mogelijke conclusie over de gehele invoerdata te verkrijgen. Het tweede deel behandelt een andere benadering waarin enkel die conclusies verkregen worden die relevant zijn voor de zoekopdrachten van de gebruikers.

In het eerste deel van het proefschrift is het MapReduce programmeermodel gebruikt om reasoning op een grootschalige manier uit te voeren en zodoende goede prestaties te behalen. In het tweede hoofdstuk wordt een reeks van op MapReduce gebaseerde reasoning-algoritmes geïntroduceerd. In het derde hoofdstuk gebruiken we hetzelfde programmeermodel om de invoerdata te comprimeren in een compactere vorm, waardoor er efficiënter gerekend kan worden met de invoerdata. In het vierde hoofdstuk gebruiken we Pig, een taal die voortbouwt op MapReduce, om grote SPARQL zoekopdrachten te coderen. Op deze manier voorzien we in een compleet systeem voor het afleiden van nieuwe informatie en het doen van zoekopdrachten, waarbij eenzelfde systeemarchitectuur en programmeermodel wordt gebruikt.

In het tweede deel verplaatsen we de aandacht naar een vorm van reasoning die wordt uitgevoerd wanneer een gebruiker een zoekopdracht uitvoerd in een kennisbank. In een dergelijke situatie kan MapReduce niet worden gebruikt aangezien de uitvoering van een dergelijke MapReduce taak een lange wachttijd oplevert. Daarom

introduceren we een nieuwe hybride reasoning-techniek waarmee we alleen een klein deel van de nieuwe informatie vooraf afleiden. Deze nieuwe informatie wordt tijdens het uitvoeren van de zoekopdracht gebruikt om de rekentijd van de zoekopdracht in te korten.

In het vijfde hoofdstuk analyseren we onze techniek van hybride reasoning vanuit een theoretisch perspectief en verifiëren we of de aanpak correct en volledig is. In het daaropvolgende hoofdstuk, hoofdstuk 6, beschrijven we een gedistribueerd en parallel prototype van deze techniek en analyseren we de prestaties op het DAS-4 cluster met een standard benchmark.

In het laatste hoofdstuk van dit proefschrift leiden we een aantal principes, die we als "wetten" beschouwen, af uit de technische bijdragen die in de vorige hoofdstukken zijn gepresenteerd. Deze wetten zijn aantoonbaar geldig voor de huidige data en zijn ook verantwoordelijk voor de resultaten die we behaald hebben in onze experimenten. De wetten kunnen worden gebruikt om een beter begrip te krijgen van de eigenschappen van het huidige onderwerp van reasoning op de schaal van het web en kunnen gebruikt worden om verder onderzoek over dit onderwerp te bevorderen.